VA MVP PHENOMIC SCIENCE

OVERVIEW & EXAMPLES

October 19, 2017

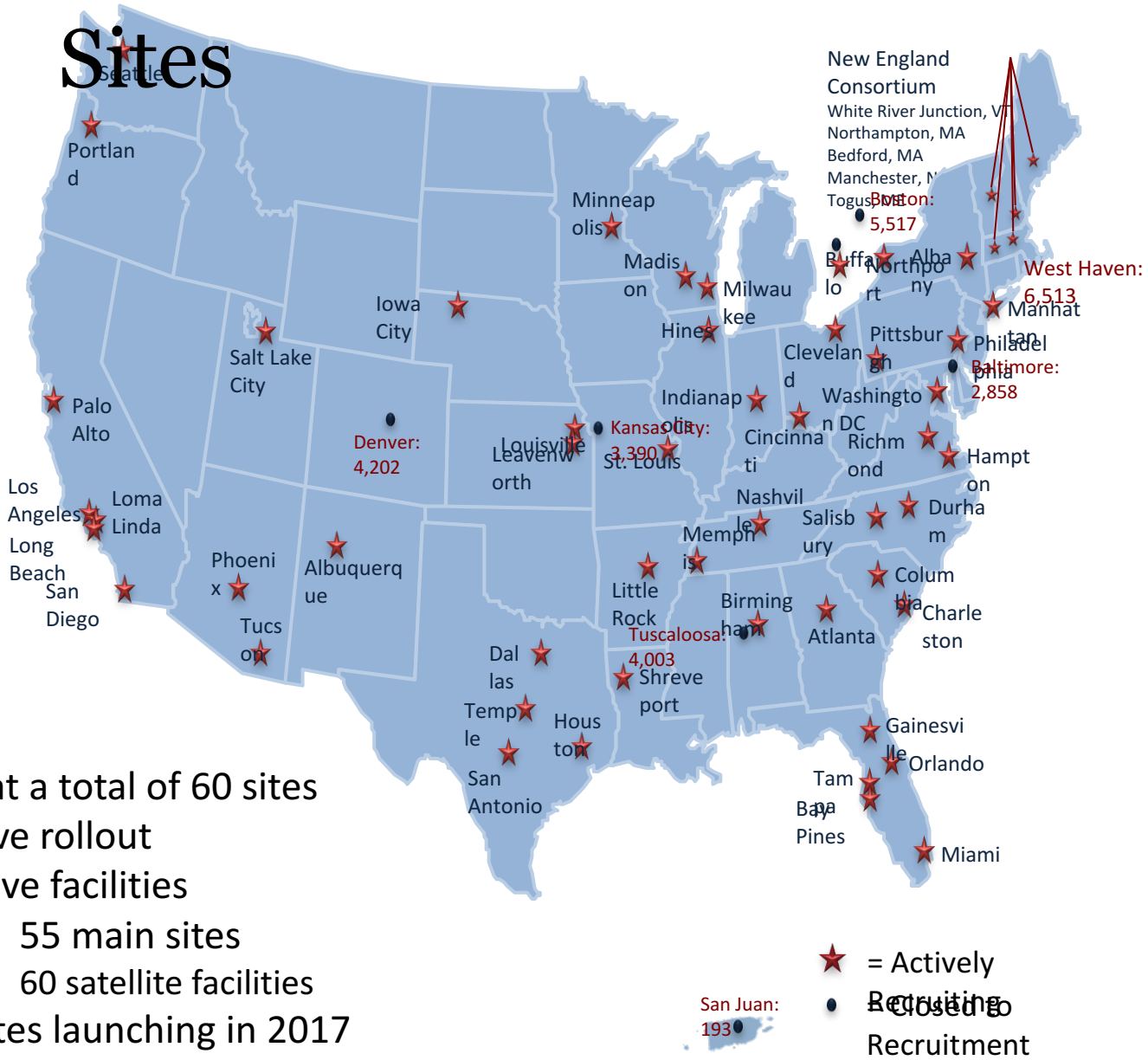
Michael Gaziano, MD MPH
Chris O'Donnell, MD
Kelly Cho, PhD, MPH
David Gagnon, MD, MPH, PhD
Katherine Liao, MD
Jackie Honerlaw, RN, MPH
Tianxi Cai, ScD

Million Veteran Program (MVP)

- Enroll up to one million users of the VHA into an observational mega-cohort
  - Collect health and lifestyle information
  - Blood collection for storage in biorepository
  - Access to electronic medical record
  - Ability to recontact participants

Million Veteran Program

MVP Enrollment Sites



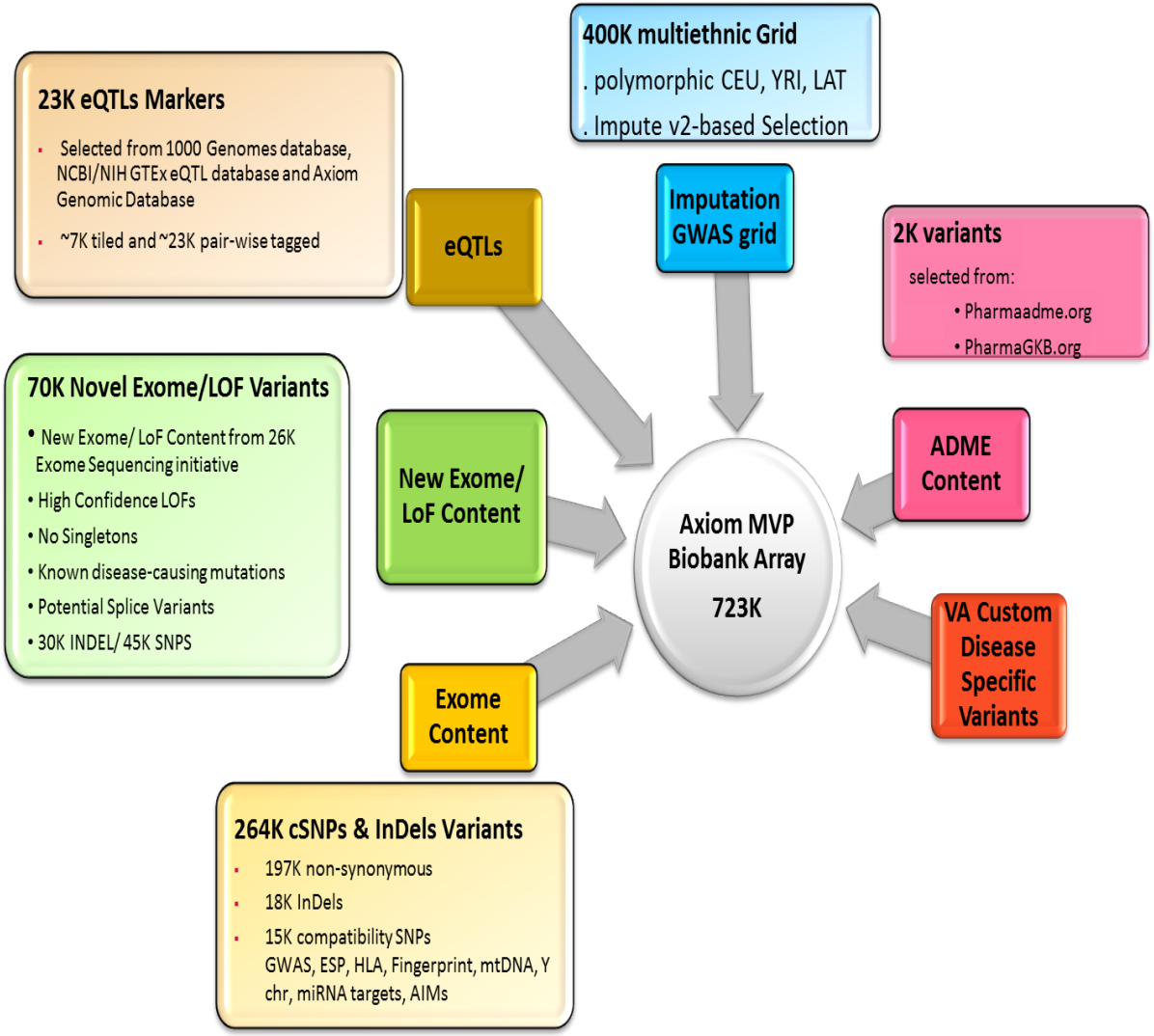
Open at a total of 60 sites

- Wave rollout
- Active facilities
  - 55 main sites
  - 60 satellite facilities
- 5 sites launching in 2017

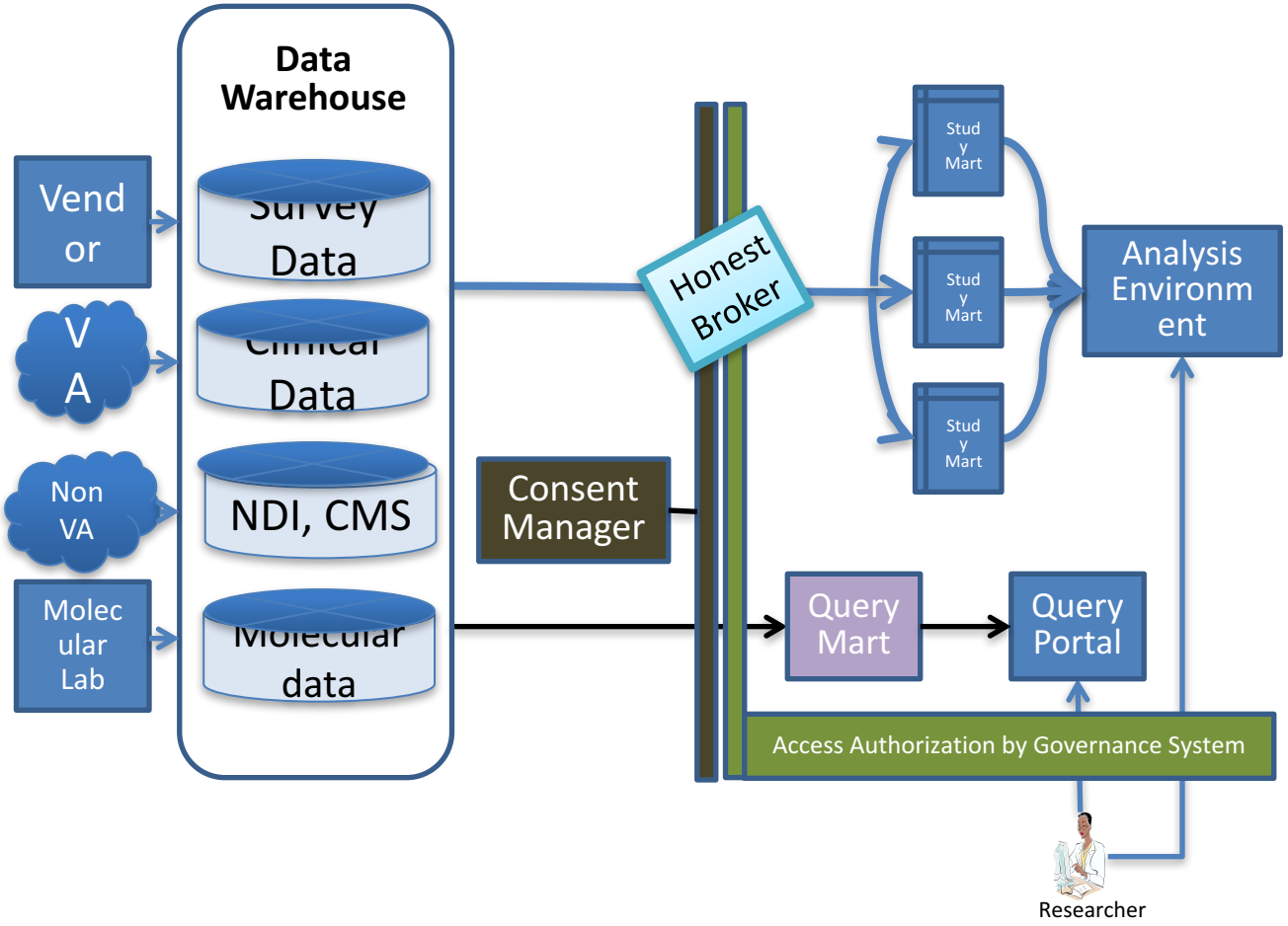
MVP Milestones

Invitation mailings sent	Over 4 Million
Consented Veterans	610,000
Completed Baseline Surveys	675,000
Genotyped, Sequenced	GT: over 500K; WGS 2K-> 45K; WES 20K
Other omics	Metabalomic, proteomic, microbiomic pilots
Funded Science	3 alpha, 5 beta, 7 gamma test projects, 3 DOE, 2 BD-STEP
Scientist, analysts on the system	80-100
Abstracts presented, submitted, preparation	7, 50, 20
Manuscripts in prep	12

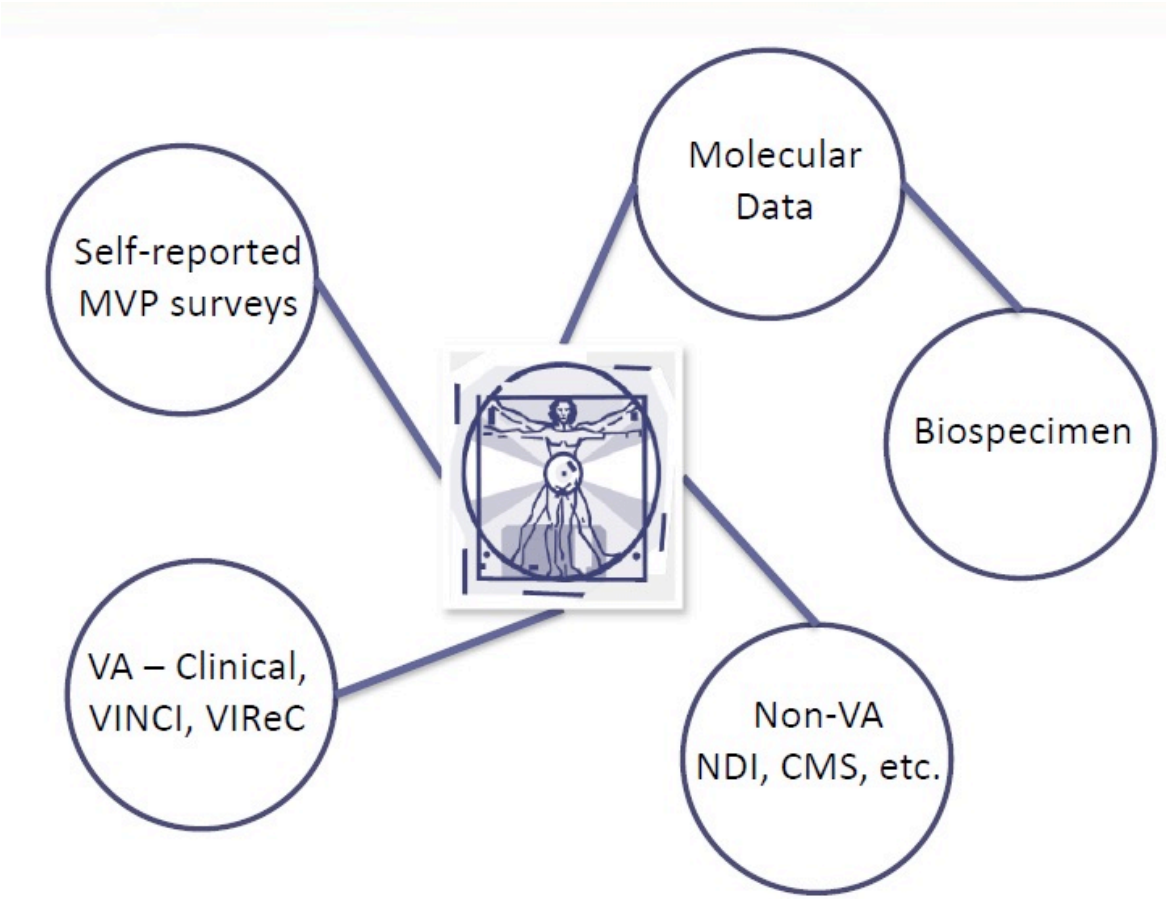
Axiom MVP Biobank Array



System Architecture

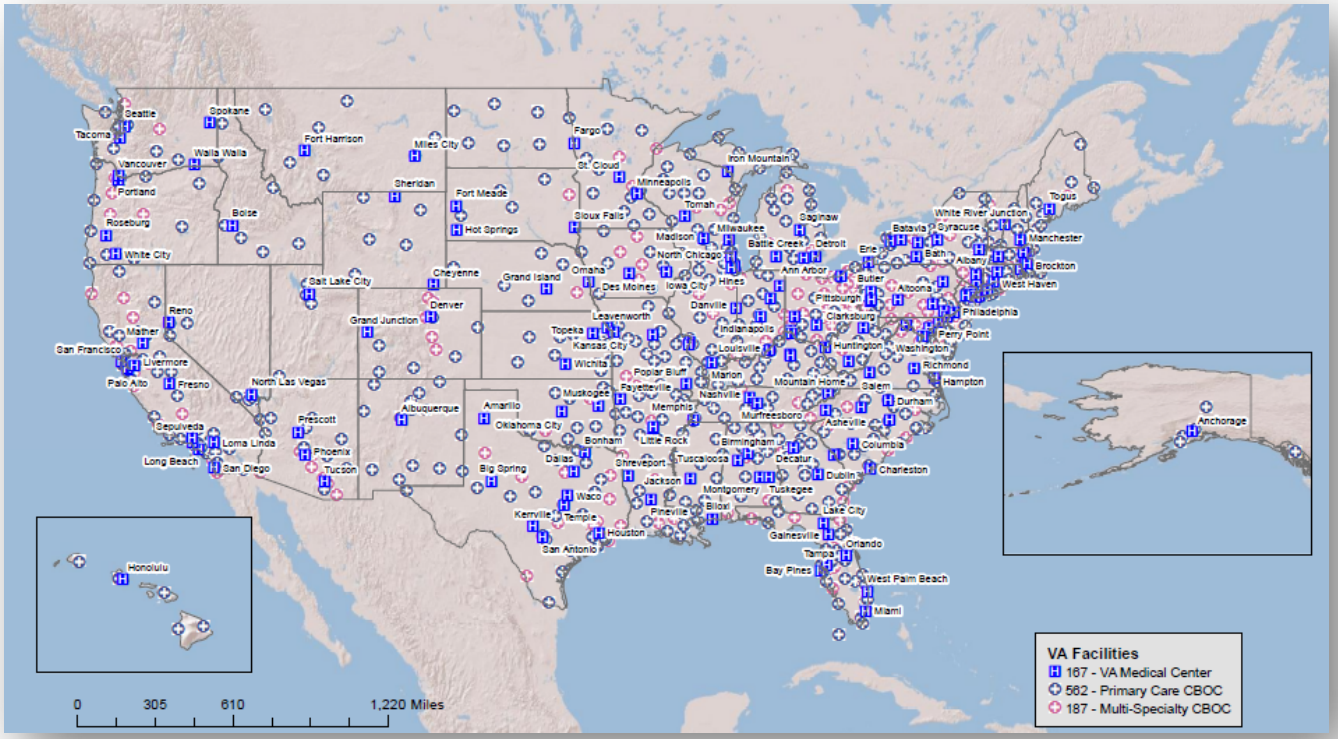


MVP Data Universe



Veterans Health Administration (VHA)

The Largest Integrated Healthcare Network in the Country



VHA Points of Care (1,748)

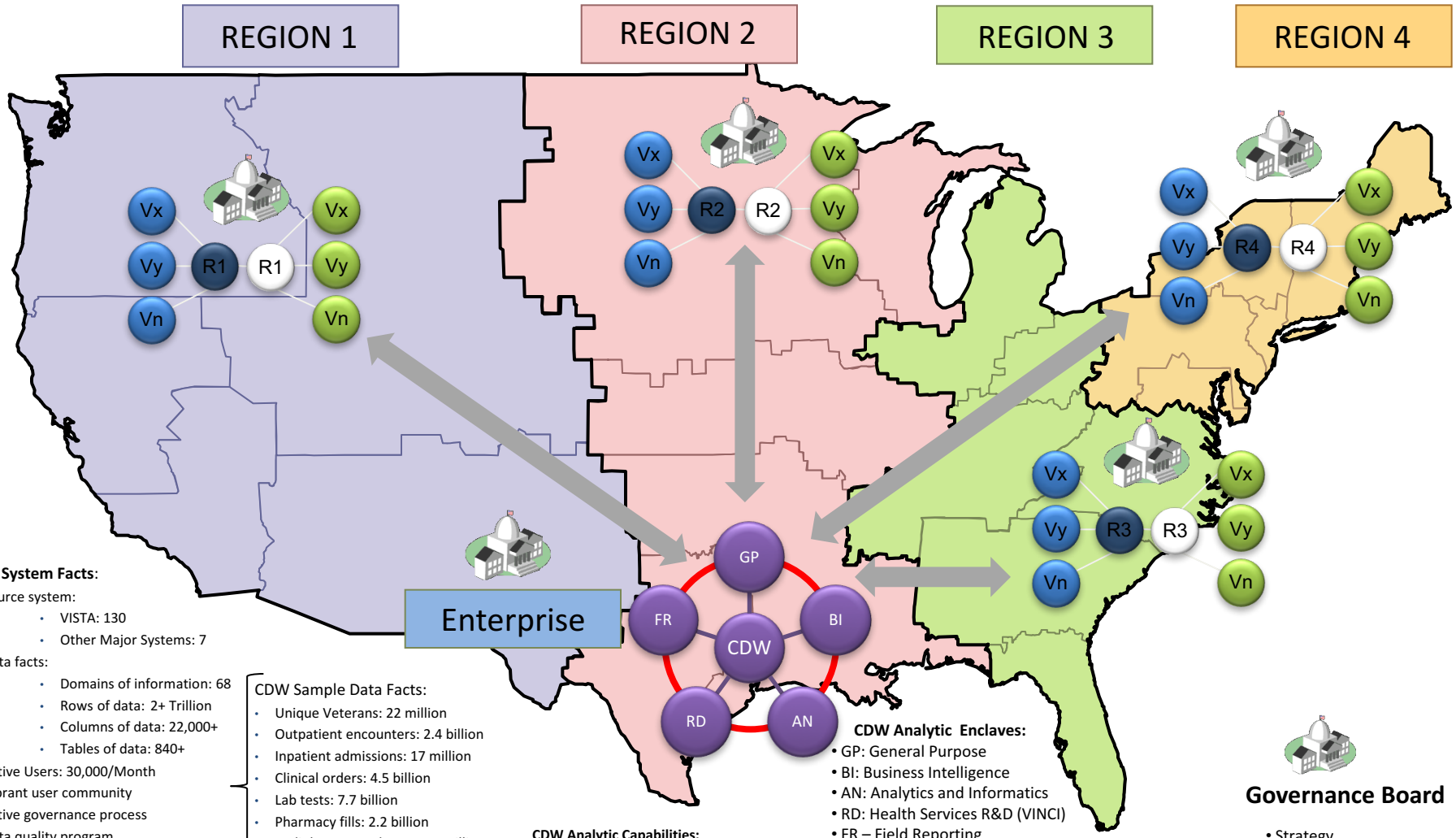
- Integrated Healthcare Networks: 21
- Major Medical Centers: 152
- Outpatient Clinics: 990
- Vet Centers: 370
- Domicillaries: 102
- Community Living Centers: 134

Patient Population

- Enrollees: 8.8M
- Active Patients: 6M
- All Time Patients: 22M
- FY15 Outpatient Visits: 84M
- FY15 Inpatient Admissions: 703K

VA Analytic Ecosystem

Common Data ♦ Common Infrastructure ♦ Common Tools ♦ Common Security



CDW System Facts:

- Source system:
  - VISTA: 130
  - Other Major Systems: 7
- Data facts:
  - Domains of information: 68
  - Rows of data: 2+ Trillion
  - Columns of data: 22,000+
  - Tables of data: 840+
- Active Users: 30,000/Month
- Vibrant user community
- Active governance process
- Data quality program

CDW Sample Data Facts:

- Unique Veterans: 22 million
- Outpatient encounters: 2.4 billion
- Inpatient admissions: 17 million
- Clinical orders: 4.5 billion
- Lab tests: 7.7 billion
- Pharmacy fills: 2.2 billion
- Radiology procedures: 202 million
- Vital signs: 3.3 billion
- Text notes: 3.2 billion

CDW Analytic Capabilities:

- Primary/Secondary/Data Mart Structures
- Data Standardization
- Metadata Services
- Business Intelligence Reporting & Dashboards Tools
- Geospatial Mapping Tools and Images
- SAS/Grid High Performance Compute Grid
- Natural Language Processing Engines
- Hadoop Cluster

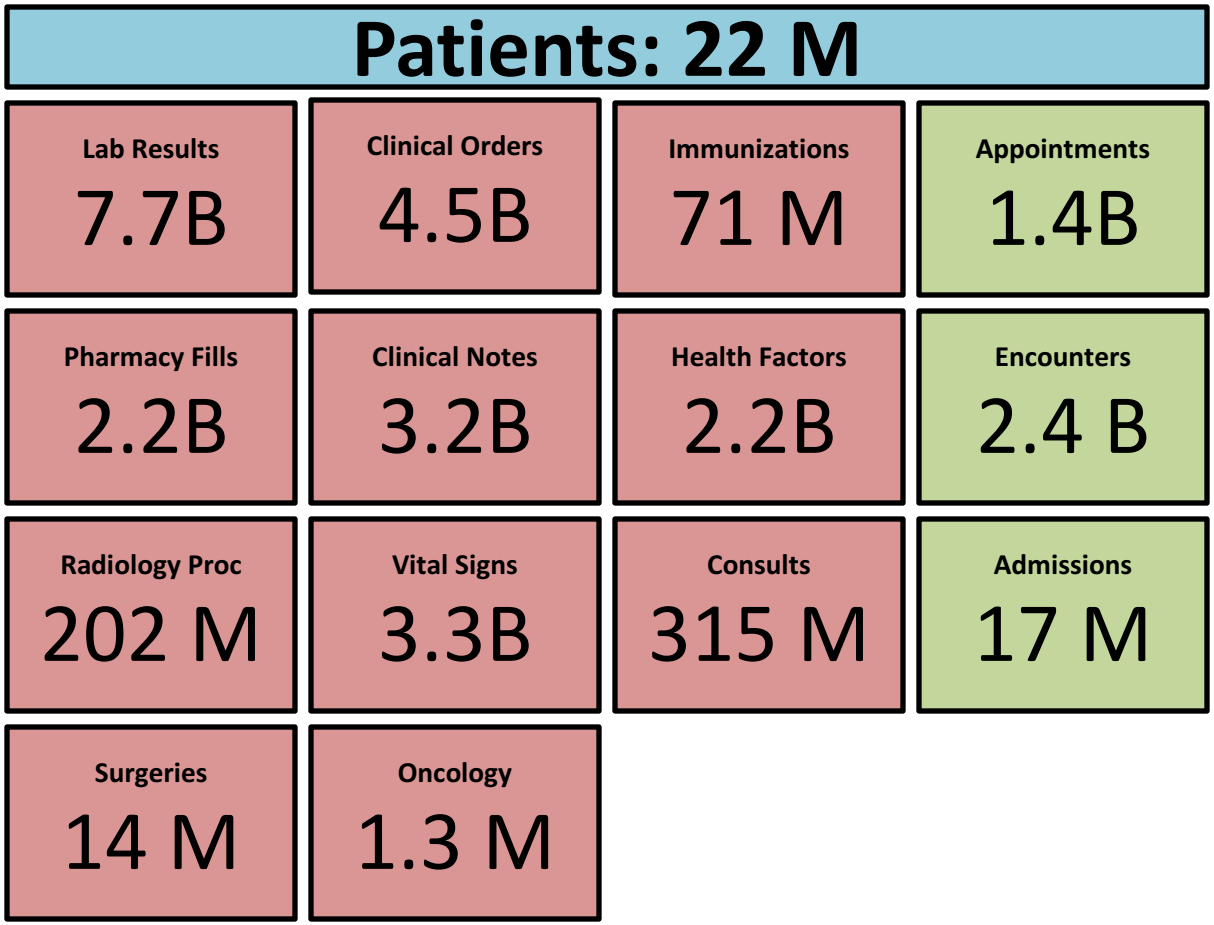
CDW Analytic Enclaves:

- GP: General Purpose
- BI: Business Intelligence
- AN: Analytics and Informatics
- RD: Health Services R&D (VINCI)
- FR – Field Reporting

Governance Board

- Strategy
- Policy
- Priorities
- Requirements

Data Examples



Domains: 15/68

VA Data Sources

- Corporate Data Warehouse Databases
- National Patient Care Databases
- Vital Status
- Decision Support System
- National Data Extract
- Beneficiary Identification Records Locator (BIRLS) death file
- New England VISN-1 Pharmacy files
- Outpatient Clinic File (OPC)
- Patient Treatment File (PTF)
- Inpatient and Outpatient Hospitalizations

National Data Systems (NDS)

- Clinic Inpatient and Outpatient Visits
- Diagnosis (ICD-9) codes
- Procedure (CPT) codes
- Pharmacy data and laboratory data

Special Data Access w/ Data Steward

- Pharmacy Benefit Management (PBM) system database
- OEF/OIF and OND Roster
- VA Clinical Assessment Reporting and Tracking (CART)
- Veterans Affairs Surgical Quality Improvement Program (VASQIP)
- Veterans Affairs Central Cancer Registry (VACCR)

General Phenotyping Approach

More and more data is becoming available for research:
is it a blessing or a curse?

- Opportunities and challenges
- Are there appropriate tools and resources to analyze, manage and handle these data?
- Are we optimally synthesizing all the information?
- Do we have all the information and annotation?

- *Sometimes, data warehouses resemble landfills more than libraries.*

MVP PHENOMICS – CORE TEAMS

Cores	Main Objectives
CORE 1: Phenomics Core Group (PCG)	○ To secure data acquisition and create Phenomics Data Universe for MVP science ○ To coordinate and facilitate phenotyping resources in support of MVP sub-studies ○ To facilitate phenotyping needs of Disease Domain Working Groups ○ To develop and maintain the MVP Phenotype Reference Library
CORE 2: Data Analytics & Management	○ To clean, curate and validate the Survey data for MVP research use ○ To maintain MVP core demographics database for analytics and reporting ○ To test and pilot Survey data elements as use cases in phenotype validation ○ To manage and organize MVP phenomics data
CORE 3: Applied Bioinformatics in Clinical Research	○ To develop methods and approaches to advance EHR data research in MVP ○ To demonstrate the application of methods to real clinical questions ○ To innovate and apply methods to solve big data phenotyping challenges

MVP PHENOMICS – CORE Tables

Table	Description
MVP Roster	List of MVP enrollees – used to create all other MVP Core Tables
MVP Baseline Survey*	MVP Baseline Survey Variables
MVP Lifestyle Survey*	MVP Lifestyle Survey Variables
MVP Core Demographics*	Standardized demographics data using CDW, OMOP and MVP Baseline Survey Data
MVP Core Vitals*	Standardized vital signs (height, weight) at the time of MVP Baseline Survey completion (uses both CDW and MVP Baseline Survey data)
MVP Core Lifestyle*	Standardized lifestyle factors (smoking status, alcohol use, exercise, nutrition scores) at the time of MVP Lifestyle Survey completion
Diagnosis Table	All ICD-9/ICD-10 codes from inpatient and outpatient encounters
Lab Table	Normalized laboratory table containing all available adjudicated laboratory tests
Medication Table	Normalized medication table containing requested VA drug classes
Vitals Table	Height, weight, blood pressure, pain score, pulse
Health Factors	Health factors related to smoking and alcohol use
CPT Procedure Table	All CPT procedure codes
ICD-9 Procedure Table	All ICD-9 procedure codes
AUDIT-C	Responses to alcohol screening survey

Laboratory Adjudication – Process

Purpose: Validate laboratory test type and results.

Example: text search for “albumin” yields 4141 tests, with only 644 that actually correspond to serum albumin – with others being, for example, urine albumin, or serum pre-albumin. Further curation is needed to identify **serum albumin**.

Adjudication Protocol	**Rationale**
1. Analyst compiles an initial spreadsheet of possible “serum albumin” tests	A text search creates an initial list of possible serum albumin tests.
2. Clinician performs initial review	Clinician reviews the name, specimen type, and descriptive statistics including total count of tests performed and average value to determine if this is indeed a serum albumin test.
3. Analyst adds relevant LOINC codes for clinician to further review	The text search may not have captured all possible serum albumin tests, so tests with relevant LOINC codes are added. (Note: LOINC codes are considered a standard but we found that they do not uniquely identify labs in the VA)
4. Second clinician performs review	Second clinician reviews, then both reviewers meet to resolve discrepancies.
5. Analyst creates final curated lab data set	The final table of accepted serum albumin tests is stored in SQL.

Examples of Laboratory Adjudication Effort

Laboratory test name	Number of tests adjudicated	Number of tests accepted
Hemoglobin A1C	527	365
Serum albumin	4141	644
Blood Glucose	4578	905
HDLC	770	377
Hemoglobin	2638	331
LDLC	1230	602
Serum Potassium	2198	720
Serum Creatinine	5212	705
Serum Sodium	2608	757
Total Cholesterol	2137	405
Triglycerides	1528	390

Serum Albumin Adjudication

Accept	LabChem TestSID	LabChem TestName	Specimen	VISN	Sta3n	Units	n	min	p1	p5	p10	p25	p50	p75	p90	p99	max
Yes	800000948	ALBUMIN(SEATTLE)	Serum	20	648	G/DL	8985	-0.22	3.1	3.7	3.9	4.2	4.4	4.6	4.8	5.2	6
No	800001031	albumin(ep), csf	Cerebral spinal fluid	20	648	%	22	51	51	54	55	57	61	66	69	71	71
No	800001092	MICROALBUMIN	Urine	20	648	MG/DL	70167	0	0.3	0.43	0.7	1.28	2.8	8.28	30.4	228.8	21321
Yes	800001119	ALBUMIN	Plasma	20	648	g/dL	712338	0.1	1.9	2.6	3.1	3.8	4.2	4.4	4.6	5	67
Yes	800001119	ALBUMIN	Serum	20	648	g/dL	21999	0.2	2.1	2.7	3.2	3.9	4.3	4.5	4.7	5.1	7.6

Medication Adjudication

Purpose: Curating VA pharmacy data requires less clinician input than adjudicating laboratory tests, but there is still considerable work required to create a usable medication dataset across data sources.

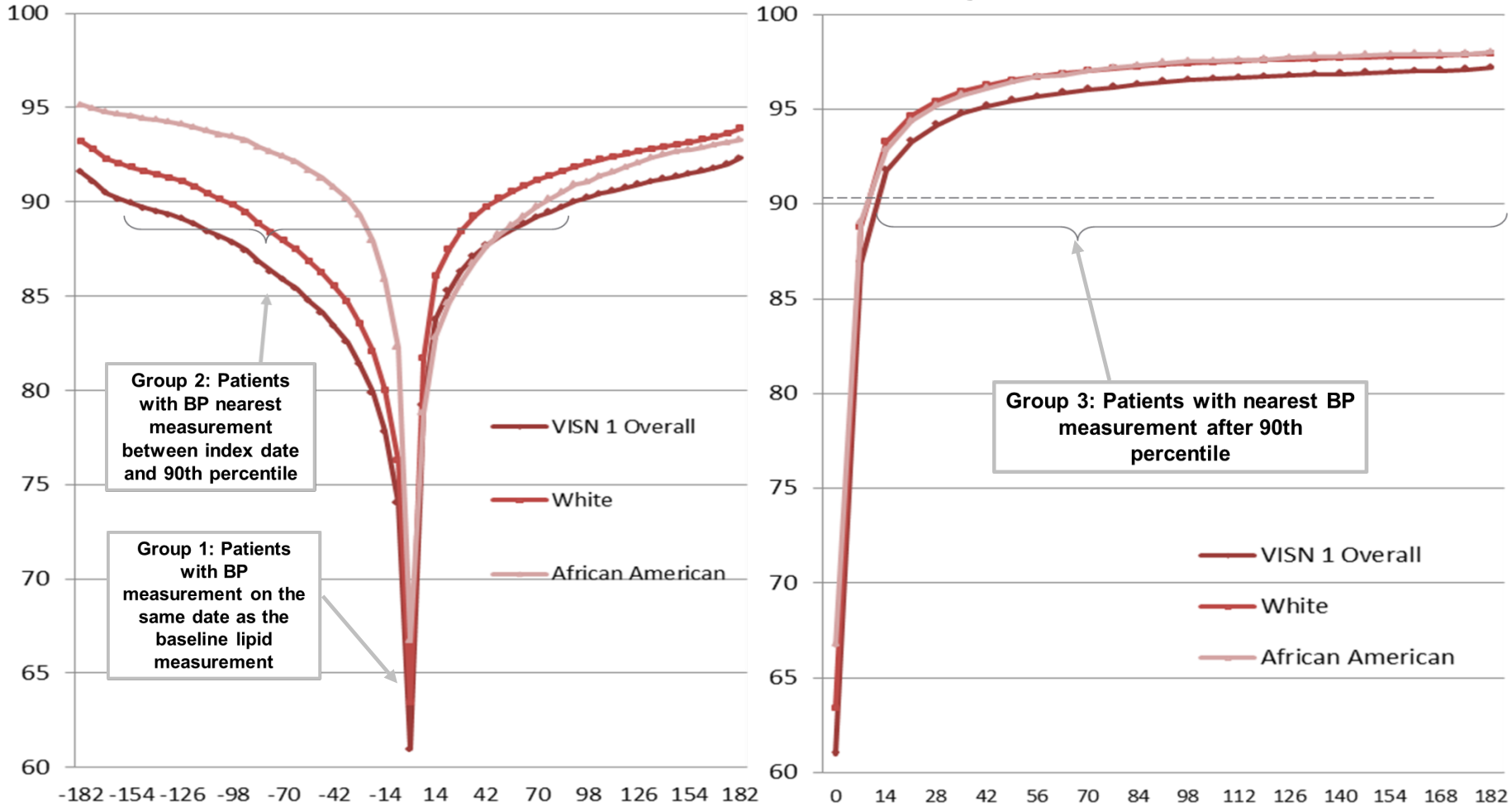
Adjudication Protocol	Rationale
1. Analyst compiles an initial spreadsheet of possible anti-lipemics	Selecting all medications in VA drug class “CV350” creates an initial list of anti-lipemics. The analyst parses out the route, dose, units and drug names from a singled field in the EMR.
2. Clinician performs initial review	Clinician reviews the list of medications and confirms if the pre-populated columns containing class, generic ingredient name, dose, units and route are correct.
3. Analyst reviews	The analyst reviews the spreadsheet to ensure that study drug or placebo drugs have not been included. Mappings to other standard naming conventions (ex: RxNorm) are incorporated into the table.
4. Analyst creates final curated lab data set	The final table of anti-lipemics is stored in SQL.

Medication Adjudication

Column	Description	Example
Variable from CDW		
LocalDrug SID	Drug ID from CDW	800170761
National DrugSID	Drug ID from CDW	800423770
LocalDrug NameWithDose	Drug name and dose from CDW	ATORVASTATIN CALCIUM 40 MG TAB
NationalDrug NameWithDose	Drug name and dose from CDW	ATORVASTATIN CALCIUM 40 MG TAB
Variable Created by Analyst		
Generic_Name1	Drug name at ingredient level – extracted from LocalDrugNameWithDose	Atorvastatin
Generic_Name2	Drug name at ingredient level, populated for combination drugs – extracted from LocalDrugNameWithDose	
Generic_Type	Sub-class – determined when identifying goal of review. In the example, the analyst is instructed to populate the subclass statin if generic name ends in -statin.	Statin
Class_Name	Class name pre-populated by analyst	Anti-lipemic agents
Dose	Medication dose – extracted from LocalDrugNameWithDose	40
Units	Medication units – extracted from LocalDrugNameWithDose	mg
Dose_Form	Route of medication – obtained from the FDA National Drug File drug table and supplemented with dose extracted from localdrugnamewithdose where missing	Tab

Class	Count	Class Name
CV050	1790	DIGITALIS GLYCOSIDES
CV100	9832	BETA BLOCKERS/RELATED
CV200	9962	CALCIUM CHANNEL BLOCKERS
CV250	6668	ANTIANGINALS
CV300	8483	ANTIARRHYTHMICS
CV350	8854	ANTILIPEMIC AGENTS
CV400	6057	ANTIHYPERTENSIVE COMBINATIONS
CV500	954	PERIPHERAL VASODILATORS
CV701	2864	THIAZIDES/RELATED DIURETICS
CV702	3468	LOOP DIURETICS
CV703	918	CARBONIC ANHYDRASE INHIBITOR DIURETICS
CV704	2431	POTASSIUM SPARING/COMBINATIONS DIURETICS
CV709	456	DIURETICS,OTHER
CV800	5499	ACE INHIBITORS
CV805	3109	ANGIOTENSIN II INHIBITOR
CV806	240	DIRECT RENIN INHIBITOR
CV900	2363	CARDIOVASCULAR AGENTS,OTHER

VISN 1 Outpatient "Virtual Baseline Data Acquisition" and Interval from Anchoring Date



Smoking Phenotype

Purpose

- To develop a probabilistic algorithm to determine smoking status of never, former, and current using CDW structured data

Gold standard smokers

- Defined using MVP self-reported smoking status from the baseline and lifestyle survey
  - 93,888 MVP year 1 genotyped participants
    - 26% never smokers; 56% former smokers; 18% current smokers

Smoking-related CDW Data (inputs)

- 1,568 smoking health factors reduced to 11 categories:

11 Health Factor Categories										
Definite Never	Definite Former	Not currently Smoking	Quit < 7 years ago	Quitting smoker	Smoker - unknown status	Definite Current	Current Chewer	Former Chewer	Chewer - unknown status	Unknown

- Smoking cessation medications
  - Bupropion HBR, Nicotine, Clonidine HCL, Bupropion HCL, Nortriptyline, Varenicline
- ICD-9/ICD-10 codes for tobacco dependence or tobacco use
- VHA clinic stop codes for smoking cessation clinic

Smoking Phenotype

Modeling

- We conducted a Least Absolute Shrinkage Selection Operator (LASSO) regression using the MVP survey response as the gold standard
- The regression coefficients were used to generate predicted probabilities of being a never, former, or current smoker
  - The category with the highest predicted probability was determined to be person's smoking status

Results

	Algorithm			
MVP Gold Standard	Never	Former	Current	
Never	19,265	4,450	427	**24,142**
Former	6,442	41,284	4,682	**52,408**
Current	322	2,163	14,853	**17,338**
Total	**26,029**	**47,897**	**19,962**	**93,888**

Never
- Sensitivity: 74%
- Specificity: 93%
- PPV: 80%

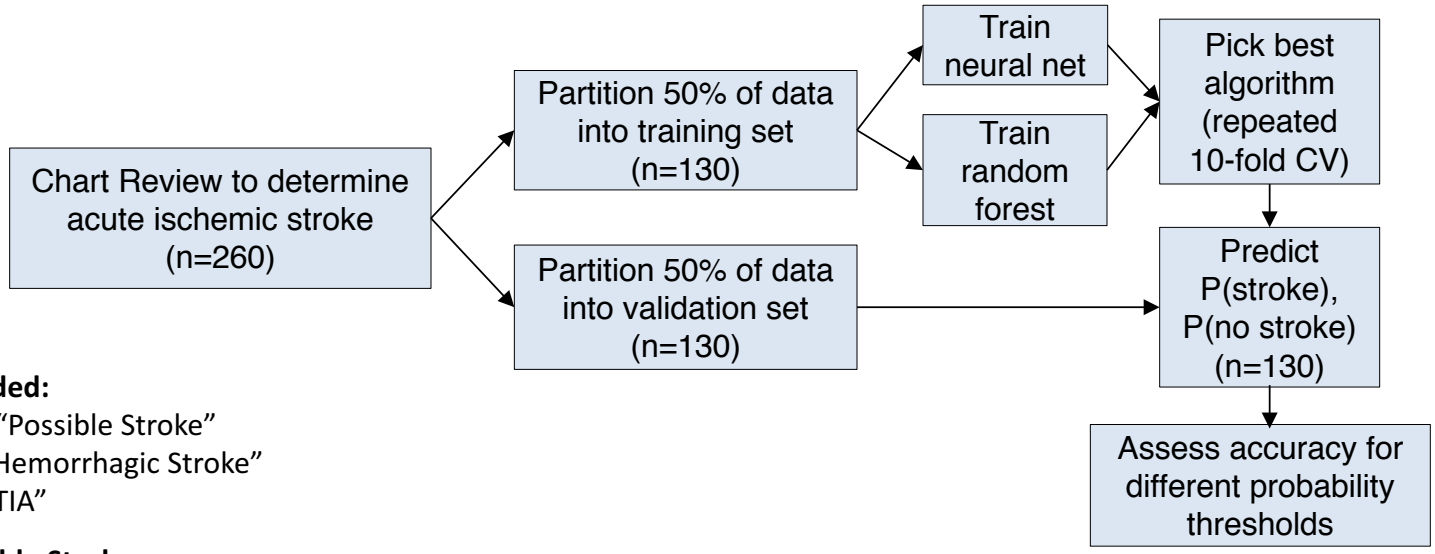
Former
- Sensitivity: 86%
- Specificity: 76%
- PPV: 79%

Current
- **Sensitivity: 74%**
- **Specificity: 97%**
- **PPV: 86%**

Stroke Phenotype - Algorithm Development

Purpose

To develop and validate a reliable protocol to identify cases of acute ischemic stroke (AIS) from a large national database.



Excluded:
n=34 “Possible Stroke”
n=3 “Hemorrhagic Stroke”
n=3 “TIA”

Possible Stroke
Relevant physician notes present, but missing primary imaging data and clinical exam at diagnosis

Stroke Phenotype - Results

Table 2. Classification Performance in the Validation Set (n=130)

Algorithm	Stroke*	No Stroke	Sensitivity	Specificity	PPV	AUC§
Tirschwell†			0.957	0.892	0.833	
Longitudinal cohort	p‡ > 0.5	p < 0.5	0.872	0.916	0.854	0.938
Case-control	p ≥ 0.85	p ≤ 0.1	0.933	0.961	0.903	0.943

* Decision rule for classifying acute ischemic stroke
† Tirschwell algorithm is Algorithm 1 from Tirschwell (2002)
‡ p is the predicted probability from the classification model.
§ Area under the ROC curve is unavailable for Tirschwell's algorithm because it is rule-based

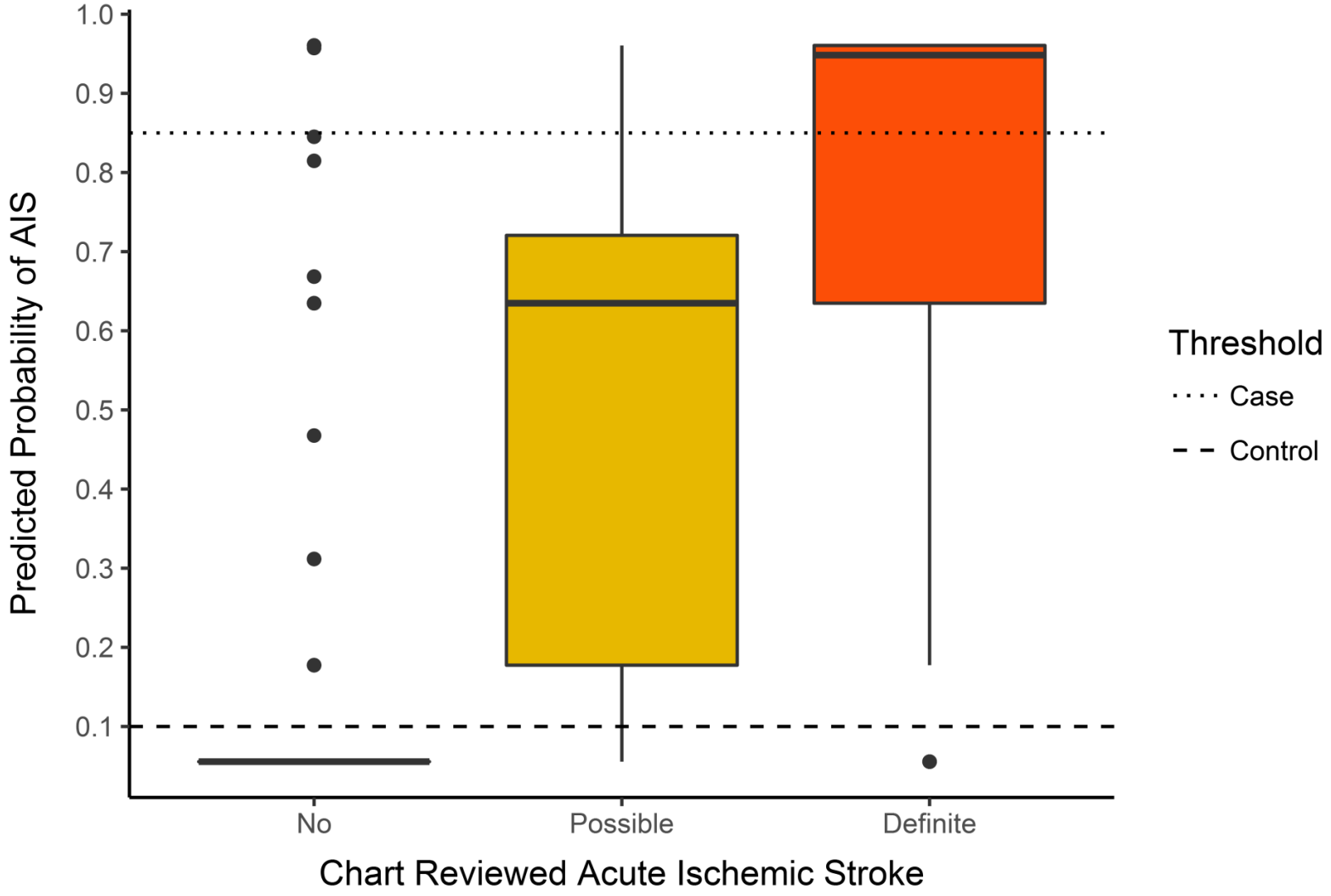
Longitudinal cohort algorithm: patient has stroke if predicted probability > 0.5

Case-control algorithm: patient has stroke if predicted probability ≥ 0.85
patient is a control if predicted probability ≤ 0.1
all other patients excluded

Case-control algorithm performs best on two fronts:
high classification metrics (sensitivity, specificity, PPV) **AND**
excludes most patients labeled as "possible AIS"
data (see boxplot on next page)

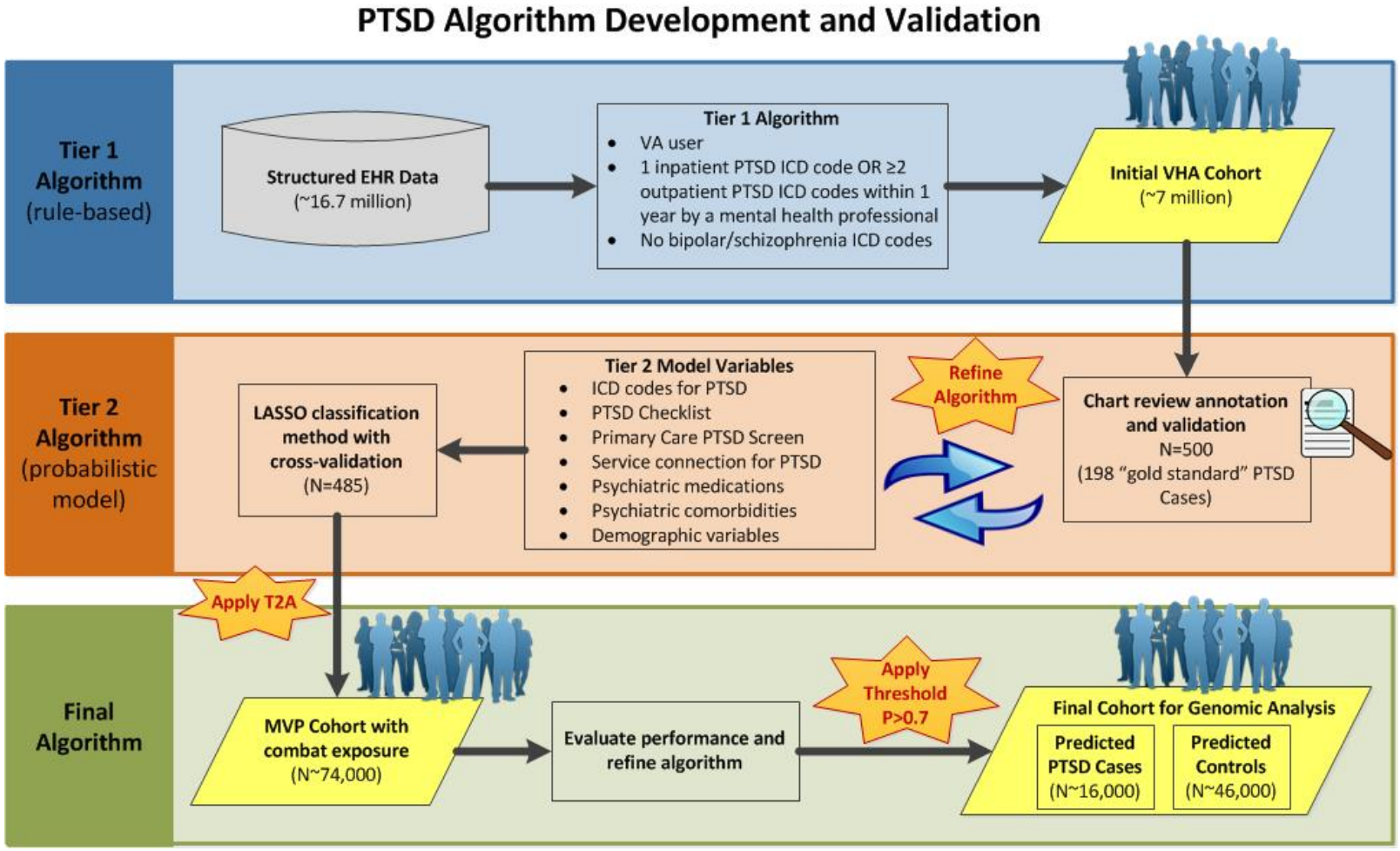
Stroke Phenotype

Case-control algorithm excludes most Possible's



Post-traumatic Stress Disorder (PTSD) Phenotype

Purpose: To develop and validate EMR-based algorithm for identifying PTSD in a sample of Veterans using a probabilistic modeling approach



This validation study was undertaken as a part of VA Cooperative Study #575B ("Genomics of Posttraumatic Stress Disorder in Veterans)," a genomewide association study of PTSD nested within the Million Veteran Program.

Performance of PTSD Algorithm

		Sensitivity* (95% CI)	Specificity * (95% CI)	PPV* (95% CI)	NPV* (95% CI)
Tier 1 Algorithm (VHA)	**Drop Possible PTSD**	1 (0.978-1)	0.995 (0.986-1)	0.961 (0.896-1)	1 (0.997-1)
	Group Possible + Case	0.877 (0.785-0.960)	0.971 (0.955-0.984)	0.792 (0.690-0.881)	0.984 (0.971-0.995)
	Group Possible + Control	0.679 (0.586-0.765)	0.979 (0.963-0.992)	0.908 (0.831-0.961)	0.912 (0.883-0.938)
Tier 2 Algorithm (VHA)	**Drop Possible PTSD**	0.995 (0.987-1)	0.995 (0.987-1)	0.995 (0.987-1)	0.995 (0.987-1)
	Group Possible + Case	0.994 (0.984-1)	0.655 (0.566-0.746)	0.907 (0.878-0.936)	0.969 (0.920-1)
	Group Possible + Control	0.951 (0.928-0.969)	0.964 (0.898-1)	0.995 (0.986-0.995)	0.712 (0.612-0.803)

* Statistics are proportionally weighted based on chart review selection

Selection of MVP Cohort for PTSD GWAS

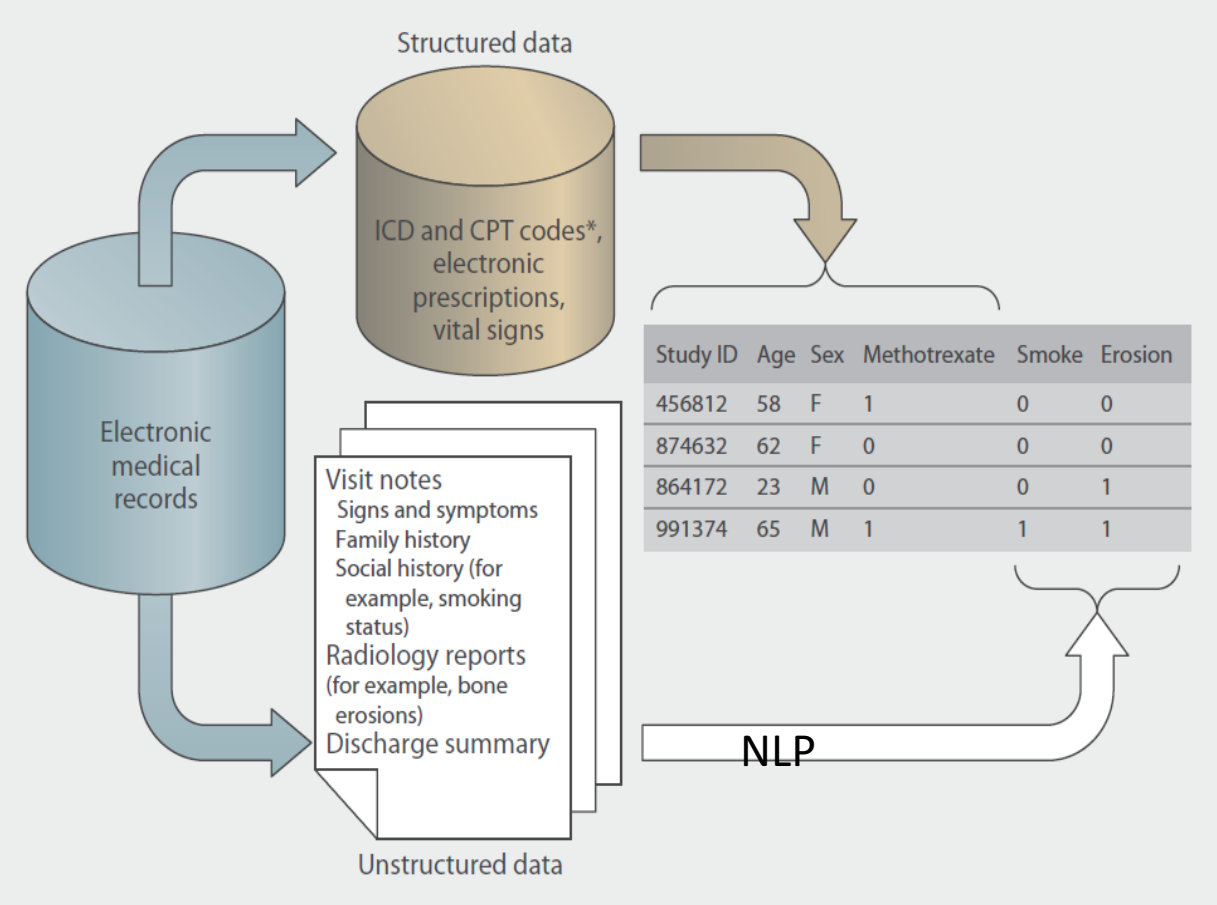
Prob(control) Cut-Off	# Controls	% Controls Retained
>0.6	48,864	97.1%
>0.7	46,319	92.0%
>0.8	38,115	75.7%

Prob(case) Cut-Off	# Cases	% Cases Retained	# Controls	Sensitivity	Specificity
LASSO	22,785	100%	46,319	0.902	0.860
>0.5	22,164	97.3%	46,319	0.907	0.858
>0.6	19,033	83.5%	46,319	0.948	0.850
>0.7	16,092	70.6%	46,319	0.977	0.837
>0.8	15,054	66.1%	46,319	0.979	0.827
>0.9	13,110	57.5%	46,319	0.984	0.809

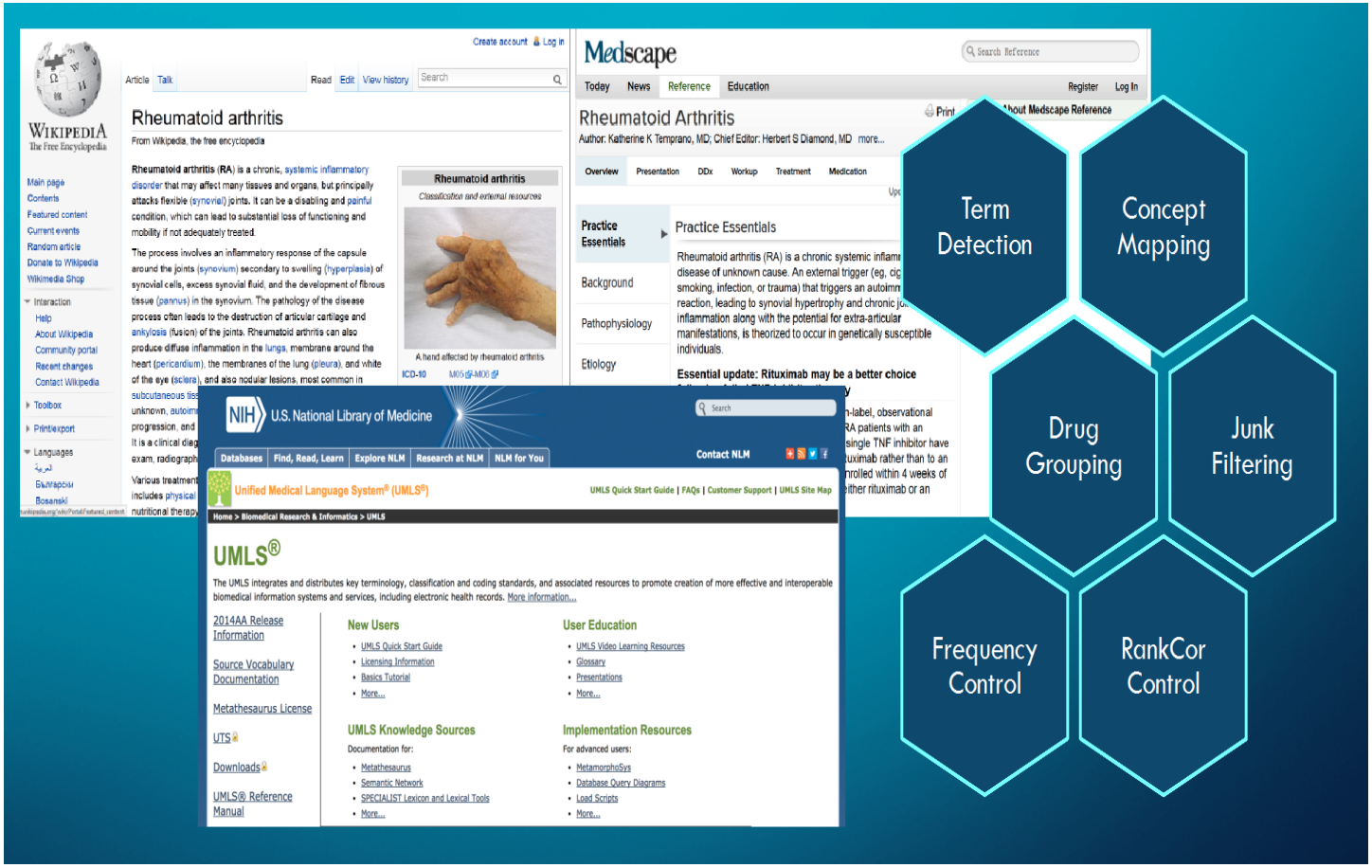
Overview: Algorithm Development and Validation Process

1) Select Initial T1 Algorithm (rules-based algorithm)
   - Based on literature review
2) Chart Validation and Evaluation of T1A
3) Build T2 Algorithm Model (probabilistic approach)
   - Literature review informed initial variable selection
   - Limited by available data
4) Iterative process undertaken to find best model for the data
5) Chart Validation and Evaluation of T2A
6) Determine Final Algorithm for GWAS (T3A)

NLP as a key component: Feature extraction

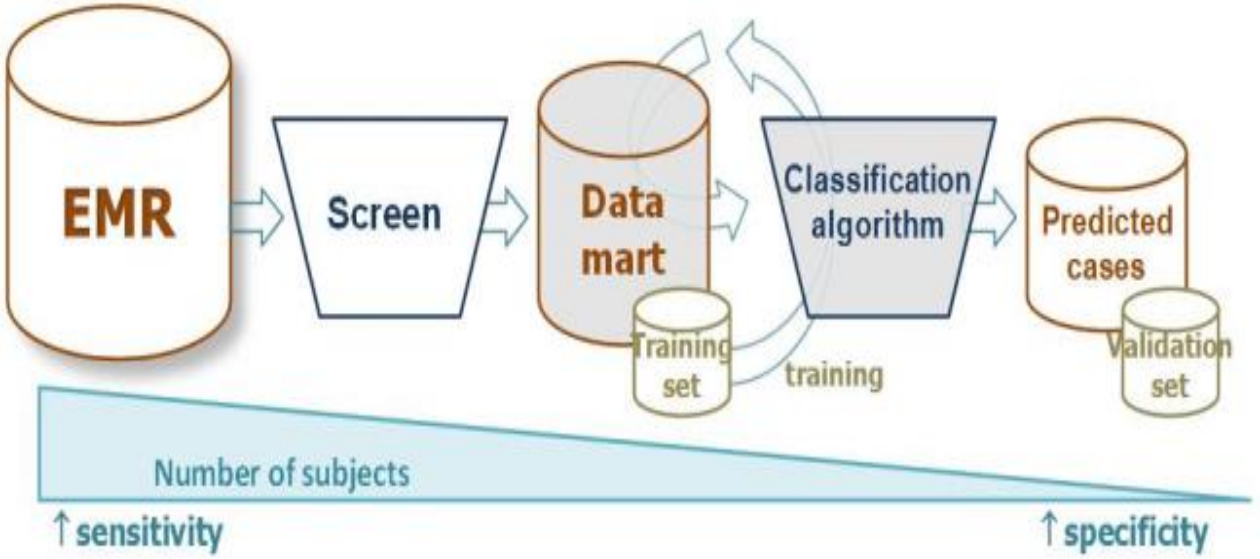


Automated Feature Extraction for Phenotyping (AFEP)



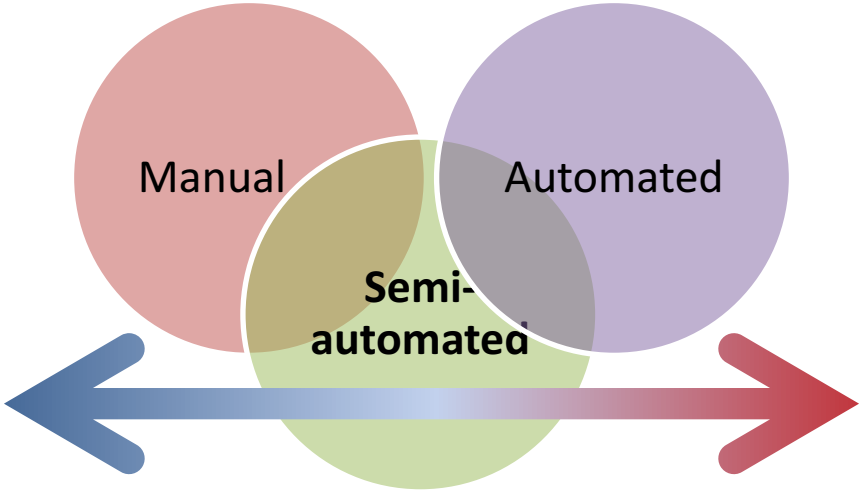
High Throughput Phenotyping Pipeline

General Framework



Cai & Liao

Our Vision for Phenotyping in MVP: A New Aproach



Semi-automated phenotyping
combines features of manual and automated phenotype development